[19]中华人民共和国国家知识产权局

[51]Int. Cl[7]
G06F 17/30

[12]发明专利申请公开说明书

[21]申请号 99805701.0

[43]公开日 2001年6月13日

[11]公开号 CN 1299488A

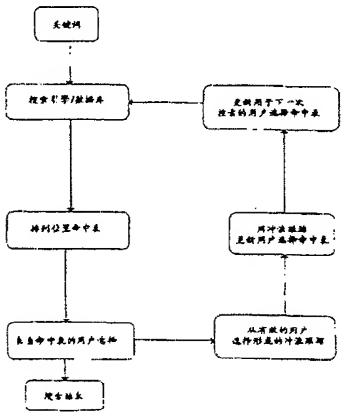
[22]申请日 1999.3.16 [21]申请号 99805701.0
[30]优先权
[32]1998.3.16 [33]US [31]60/078,199
[32]1998.7.15 [33]US [31]09/115,802
[86]国际申请 PCT/US99/05588 1999.3.16
[87]国际公布 WO99/48028 英 1999.9.23
[85]进入国家阶段日期 2000.10.31
[71]申请人 NBCI新西兰有限责任合伙公司
地址 美国加利福尼亚
[72]发明人 格兰特·J·瑞恩 肖恩·W·瑞恩
克雷格·M·瑞恩 韦恩·A·芒罗
黛尔·鲁宾逊
[74]专利代理机构 中国国际贸易促进委员会专利商标事务所
代理人 于 静

权利要求书11页 说明书52页 附图页数27页

[54]发明名称 改进的搜索引擎

[57]摘要

本发明提供一种利用用户对特定网页列表的选择结果更新因特 网搜索引擎数据库的方法,该特定网页列表的选择结果来自提供给 该用户作为其初始关键词搜索输入的结果的一般网页列表。通过用 许多不同用户的选择来更新该数据库,可更新该数据库以便按重点 排列相对给定关键词选择最多的那些网页列表,从而在以后使用相 同关键词搜索输入的搜索中首先给出最流行的网页。



ISSN1008-4274

关键词建议者：这是由用户已找到的有用的，从连续冲浪者关键词表编译的其它关键词的永久排列的数据集，并链接到每个关键词(这等同于该用户的选择的命中表)。

输入数据集	输出数据集
·冲浪者关键词表(临时和永久) ·现有用户的选择的命中表(永久)	·新用户的选择的关键词表(永久)

用户基本搜索算法

上面提供的讨论提供了更全面地描述本发明所需的语言。如图3A和3B所示，该图提供了根据本发明的搜索引擎能力的概要，其中在形成提供给终端用户的搜索结果的过程中选择网页。在步骤112，用户输入多达4组数据：关键词52，简介类型54，搜索类型58和用户ID56。IP地址62和日期时间60不由用户输入，但用户使用该搜索引擎时可读出。在步骤114和116中并行使用该数据以产生网页列表。下文详细讨论的步骤114是从根据本发明产生的新颖的新搜索引擎数据集选择网页的过程。如果需要，这可与步骤116并行运行，步骤116获得从其它现有搜索引擎选择的网页。此后，组合从步骤114和116选择的网页并在步骤118标记。下面更详细描述的标记网页列表的处理可生成图3中冲浪者跟踪数据所示的数据集并在搜索引擎用户从步骤120中的列表选择网页时发回到搜索引擎。选择标记号的网页的处理生成后面的数据系列，用该数据系列更新搜索引擎数据集：关键词124，URL126，用户ID130，日期时间132，简要网页说明134。

虽然最好是使用冲浪者跟踪数据中的所有这些不同的数据类型，使用该数据的不同组合完全在本发明的要求范围内。当新网点加到搜索引擎10的数据集114时，描述134通常仅包括在本发明的优选实施例中，所使用的描述将是出现在原始网页列表上的描述。如

下面进一步说明的，日期时间数据132可仅表示一个网点被选择，而不是记录用户在特定网点的时间周期。在从网页列表选择网页时，该过程是在步骤122直接占用对应的URL的用户看不见的。下面更详细地描述步骤114，118和120的实施细节。

在初始选择之后，用户可选择访问另一个网页URL搜索结果。根据该网点的关联，用户可以费时读取，下载，探索进一步的网页，嵌入的链接等时间，或如果对该网点的出现不关联/不感兴趣，在短周期后用户可以直接返回该搜索结果。记录两个选择之间的时间差作为与来自网页搜索表的后续选择的两个日期/时间数据132之间的差(在该实施例中，如果在访问该网页后做出另一个选择，人们可仅测量在一个网页花费的时间，然后将其提供给允许计算时间差的另一个冲浪者跟踪132)。有关该网页流行性的冲浪者跟踪数据被用来排列后面的搜索，如下文进一步描述的。

因此，根据本发明，推论和分析将其用来建立与该搜索主题的不同结果的关联是人类用户的能力。本发明采用所有人类用户的累积处理和推论，以便提供比上面可能的方法类型获得所需信息源有效得多的手段。

如上所述，通过在每个关键词搜索后记录用户去了哪个网页来捕捉人脑的力量。根据本发明，通过在对用户搜索产生网页列表中把自动发送信息的隐藏链接发回到搜索引擎(或辅助服务器)实现对冲浪者跟踪数据的集中。虽然用户仅看到显示了其要求的链接，隐藏的链接通知可用Java小应用程序执行处理的转移的搜索引擎。因此，当因特网用户选择一个网页时，它将用户引向该地址，但是还向搜索引擎10发出指出已选择了什么的冲浪者跟踪数据。当用户返回到网页列表和选择另一个网页列表时，则执行生成另一个冲浪者跟踪的另一个Java小应用程序。来自两个后续选择的该冲浪者跟踪中的数据时间数据之间的差别捕捉该用户已在前一个网站的时间周期。用户不了解该数据被发出的情况发生。

在另一个实施例中，不是使用多个Java小应用程序集中冲浪者

跟踪数据的完整列表，没有描述数据134，日期时间数据132表示一个用户访问了一个特定的网站。在一个具体的实施例中，用户必须访问特定的网站比预定的时间周期多，例如1分钟或15分钟，取决于什么是已查看该网点以便对访问的网点计数和向搜索引擎10发回任何冲浪者跟踪数据的适当时间，同样如下文所述。在该实施例中，每个小应用程序包含在搜索引擎更新数据库所需的所有信息。另一个实施例在用户导航到要求的网站之前集中冲浪者跟踪数据。获得该冲浪者跟踪数据的其它方式也是可能的并且在本发明所要求的范围之内。

因此，根据本发明的搜索结果页是与常规的搜索引擎的结果页不同形成的。其差别在于作用而不是内容。在视觉上，对用户来说，该页看起来与来自其它搜索引擎的标准搜索结果相同。

一个实例说明这一点：在常规搜索中，对关键词"Weather"搜索的结果页可以读出：1.www.weather.com 今天的天气预报。预期今天各地晴好。

与"www.weather.com"标记相关联的HTTP链接是"http://www.weather.com"。这表明：如果用户选择该链接，直接将他们导航到该网页。

相反，根据本发明，使用关键词"Weather"进行的搜索的标记的结果网页可以读出：1.www.weather.com 今天的天气预报。预计今天各地晴好。

与"www.weather.com"标记相关联的HTTP链接是：link.asp?n=1."。因此，如果用户选择该链接，在用户看不见的处理中，根据本发明首先将用户引导到与搜索引擎10的网络服务器对应的网点上的link.asp网页，并传送值为1的参数n。

服务器侧的代码(在万维网服务器上运行的应用代码)使用该参数识别URL和用户的选择网点的描述。然后将该信息与其它冲浪者跟踪数据一起存储在数据库表中。然后服务器侧代码对用户需要的URL执行改向操作。用户则看到其所需的网页出现。

搜索结果的来源与该活动无关。用户的目的地网页与该活动无关。该处理是将用户，关键词和目的地记录到数据库中的一种。该跟踪方法可仅记录关键词搜索后最初访问的网页。如果用户连续返回到该搜索结果表，则可记录后面的网页访问。

可对作为用户选择网页的结果发回到搜索引擎10的数据集114的冲浪者跟踪数据加密，以防止欺诈的用户把假数据发回到搜索引擎。

用户可从最初的URL选择连接到(如果他们不返回搜索结果页)的另一种跟踪方法是运行所选择的网页作为位于搜索引擎网站的'帧'的部分。这样允许在输入关键词后记录所访问的网页的完整记录。然而，这样会另外增加系统的复杂程度，可能会降低系统响应时间。

如前面提到的，由于其提供了搜索者(冲浪者)如何使用搜索引擎的跟踪或记录，可收集的冲浪者跟踪数据包括关键词124，URL126，用户ID128，IP地址130，日期时间132，简要网页说明134，并这样来识别。该数据用来改善在前面的搜索者的喜好上建立的未来的搜索。因此，冲浪者跟踪是单独用户或万维网'冲浪者'针对特定的关键词组从初始搜索结果优选的一种手段。如何生成确定网页列表的数据集合。

图4说明了根据本发明用于不同网页搜索的数据集。用于确定该网页列表的数据集包括关键词表164，简介ID表166，安全表168，累积的冲浪者跟踪表170，关键词URL链接表172，个人链接表174，和网页(URL)表188。

下面更详细地描述上述数据集的结构。后面的描述给出以电子表格方式排列的数据，每个单元有多个值并有多个空白单元。以这种方式进行说明便于解释本发明，但不是一种有效的存储和检索方法。关系数据库模型用于根据本发明实现数据存储，以致有多个所涉及的字段或表来存储数据对于本领域的那些技术人员来说是很明显的，每个字段仅存储一个值。

关键词表(164)

在下面所示的表1中更详细地给出图4的关键词数据表164的内容，并且是关键词表，包括短语，和已向他们请求的次数。如果该表变得较大以致不能管理，可从该列表删除预定时间周期之后不再使用的关键词。然而，如果可能，希望保留输入的大部分或所有关键词短语。

关键词	该关键词被请求的累积次数(W)	每个关键词的唯一编号
关键词1	W1，W2，W3等	
关键词2		
关键词3		
关键词4		
关键词5		
关键词6		
关键词7		

表1 信息请求列表和其被请求的次数

可根据所选择的(W1，W2，W3，...)，例如W1=总搜索，W2=男性简介，W3=女性简介，W4=USA简介等不同的"用户简介"来分离一个关键词被请求的累积次数。应指出，由于用户可落入多于一个简介类别，例如来自USA(W3)的男性(W2)，W的总和比一个网点已被访问的总次数大。这不仅将变成使用该关键词的用户搜索者的数量，而且是用户(根据所选择的简介类型)搜索该关键词的类型表。如下文所述，虽然他们涉及使用关键词建议器，只要拼写不同，表明相同事物的关键词在不同语言中是不同的关键词。

网页表(188)

下面所示的表2中更详细地给出了图4的网页表188的内容，并包含因特网网页列表。每个网页具有URL地址，相关联的2-3行说明，对每个URL(也可以是任何字符，符号代码或表达式)唯一的网页数量，和该URL已被访问的累积次数。URL地址具有向其分配的唯一

编号(也可以是任何字符，符号代码或表达式)，而不是在后面的数据表中存储的完全URL串。

地址	2-3行说明	每个URL地址的唯一编号	URL(网页)被访问的频率
URL地址1			
URL地址2			
URL地址3			
URL地址4			
URL地址5			
URL地址6			
URL地址7...			

表2是信息提供者列表和网页说明

关键词URL链接表(172)

下面所示的表3中详细给出了图4的关键词URL链接表172的内容。由于该表包含信息提供(URL地址或网页)和信息请求(关键词)之间有关链接的信息，它对于本发明来说特别重要。

该数据将来记录在描述关键词与由下面的三个参数定义的具体值之间的关系的数据集。

-对与每个关键词对应的每个URL地址的有效访问(命中)的累积数量(在此称之为X或加权系数X)。这是每个关键词的URL的流行性的量度并从冲浪者跟踪确定。

-在更早的预定时刻测量的有效访问的以前的累积数量(在此称之为Y或加权系数Y)

-与每个所述网页的生成或输入时刻有关的日期时间系数(在此称之为Z或加权系数Z)。Z是网页开发者向搜索引擎提交网页的日期时间。

并不是关键词和URL地址的所有组合都具有数据X，Y和Z。

	关键词	关键词	关键词	关键词	关键词
URL地址1	X,Y,Z				
URL地址2					X,Y,Z
URL地址3			X,Y,Z		
URL地址4	X,Y,Z				
URL地址5		X,Y,Z		X,Y,Z	
URL地址6					
URL地址7					

表3是信息提供者(网页)与信息请求(关键词)之间的链接

具有该关键词URL链接表的简介类型

网页的流行对于不同的人群是不同的。包括多种简介类型将产生表3中的多个X，Y和Z的值，例如人们可具有由X1 X2 Y1 Y2等表示的全球和新西兰的流行速度。

	关键词"体育"
与橄榄球有关的URL地址	X1=520，X2=52
与篮球有关的URL地址	X1=4000 X2=20

在该例中，橄榄球和篮球URL地址的全球流行性(使用一般的简介类型)分别是520和4000，对于新西兰的简介类型分别是52和20.

当使用一般的简介类型设定时(根据X1排列)，篮球网点排列在顶部。当选择新西兰设定时(根据X2排列)，橄榄球网点将是最高的。这反映了新西兰人的喜好。这是一种存储不同人群的喜好的非常简单的方法。

人们会期望基于新西兰的橄榄球网点在新西兰表上比海外网点评定高，但没有必须是这种情况的原因。西班牙在世界上可能具有最好的橄榄球网点。该系统仅对用户感觉的信息质量评估网页，网点的物理位置并不重要。

表示不同国家，职业，性别，年龄等的X值有非常大的范围，能够非常简单地捕捉不同人群的流行性。用户可根据其个人的兴趣/特征选择组合任何X值。

作为例子，如果说，

· X1表示男性

· X2表示女性

· X3表示新西兰人

· X4表示美国

· X5表示工程师

· X6表示律师

一个"男性"和"新西兰人"将使用X3和X1两个搜索引擎增量。该便利条件增加了系统的数据需求，但它对不同的用户将极大地改善搜索结果。由于用户可加入一种以上的人群，因为用户可属于一种以上的简介类型，需要将网页的总流行率作为分开数量存储。所有单独流行率之和将大于总流行率。

为针对用户简化该系统，存在一种缺省简介类型(X的选择)，该缺省简介类型具有一个使用其它简介类型进行特定搜索的选项。例如，用户可具有新西兰男性的缺省简介类型，但如果需要一种技术，可选择一个反映全世界工程师的累加的搜索知识的'全球工程师"简介类型。

人格化的程度取决于搜索的频率。例如，诸如"新闻"之类的常用关键词将具有高人格化程度(大范围的X值)，如"英语邮票"之类较不太普遍的关键词有很少或没有人格化(仅有全球X值)。人格化的程度是使用该关键词(从表1找到的)的频率的函数。

累积冲浪者跟踪表(170)

下面所示的表4更详细地给出图4的累积冲浪者跟踪表170的内容。用冲浪者跟踪数据更新与网页和表3中的关键词(也称为关键词URL链接表172)之间的链接有关的信息。累积的冲浪者跟踪是从所有独立的冲浪者跟踪组合的信息，用它来确定"命中"多少(有效访

问)，每个网页针对每个关键词。

从每个独立的冲浪者跟踪收集的信息是前面描述的一串输入，并在下面以表格的形式给出

IP号码	用户ID	关键词	URL(网页)	日期时间

表4中每一行是一个冲浪者跟踪，组合的行是累积的冲浪者跟踪

下面进一步处理冲浪者跟踪数据以更新表3的方式

简介ID表(166)

下面所示的表5中更详细地给出图4的简介ID表166的内容．该表包括唯一标识，密码，联系电子邮件和他们通常用来进行搜索的缺省简介类型。

用户标识	密码	电子邮件	缺省简介	其它信息
Joe Bloggs	dogs	jbloggs@AOL	US，男性	

表5是用户识别表

存储用户缺省简介类型作为用户的个人喜好简介部分，通过向系统输入某些个人识别的形式来访问该简介。当登录到数据搜索引擎可提供该信息，或如同本领域中已知的项目，搜索引擎可在计算机上留下"cookie"以识别用户，(应有一个任选的电子邮件地址和与登录过程相关联的密码(或类似内容)。IP地址本身不是充分的识别手段，由于它对各个用户来说不一定是唯一的。

其他信息可以包括用户定义的喜好如何对搜索结果进行组合以及特定用户感兴趣的关键词。该信息可用于主动地定制访问网页的搜索结果和建议。

个人链接表(174)

下面所示的表6中更详细地给出图4的个人链接表174的内容。表

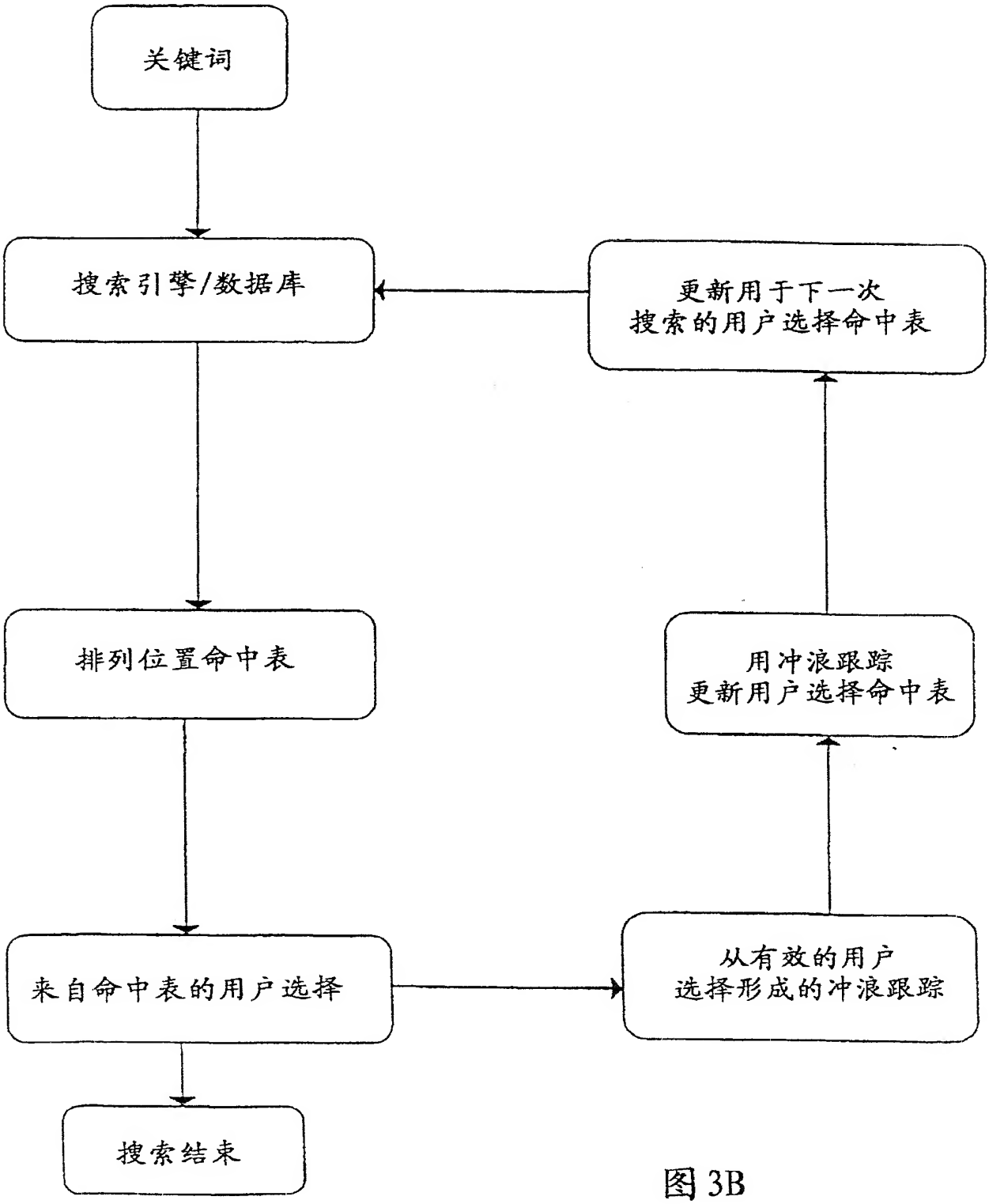

关键词
搜索引擎/数据库
更新用于下一次
搜索的用户选择命中表
排列位置命中表
用冲浪跟踪
更新用户选择命中表
来自命中表的用户选择
从有效的用户
选择形成的冲浪跟踪
搜索结束

图 3B